J Mol Hist (2009) 40:343–352
DOI 10.1007/s10735-009-9247-9

ORIGINAL PAPER

Comparative analysis of an experimental subcellular protein localization assay and *in silico* prediction methods

Yuhui Hu · Hans Lehrach · Michal Janitz

Received: 30 September 2009 / Accepted: 1 December 2009 / Published online: 22 December 2009

Abstract The subcellular localization of a protein can provide important information about its function within the cell. As eukaryotic cells and particularly mammalian cells are characterized by a high degree of compartmentalization, most protein activities can be assigned to particular cellular compartments. The categorization of proteins by their subcellular localization is therefore one of the essential goals of the functional annotation of the human genome. We previously performed a subcellular localization screen of 52 proteins encoded on human chromosome 21. In the current study, we compared the experimental localization data to the *in silico* results generated by nine leading software packages with different prediction resolutions. The comparison revealed striking differences between the programs in the accuracy of their subcellular protein localization predictions. Our results strongly suggest that the recently developed predictors utilizing multiple prediction methods tend to provide significantly better performance over purely sequence-based or homology-based predictions.

Keywords Protein localization · *In silico* predictions · Human chromosome 21 · Immunocytochemistry

Y. Hu · H. Lehrach · M. Janitz
Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany

Y. Hu
Max Delbrück Center for Molecular Medicine (MDC) in der Helmholtz-Gemeinschaft, The Berlin Institute for Medical Systems Biology, 13125 Berlin-Buch, Germany

M. Janitz (✉)
School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia
e-mail: m.janitz@unsw.edu.au

Introduction

Knowing the location of a protein within its cellular environment is critical for understanding the regulatory mechanisms by which it is controlled. The accurate function of proteins and their interaction networks relies greatly on the proper localization of each protein component. A conventional method to identify protein–protein interactions at the single-cell level is to trace the mutual localization of proteins under physiological conditions (Relic et al. 1998; Surapureddi et al. 2000). Another common strategy in the study of regulation and interaction networks is to determine whether the localization of proteins is altered by the intentional disruption of the networks (Zuckerbraun et al. 2003). The aberrant translocation of proteins often correlates with pathological changes in cell physiology and accounts for the clinical manifestations of several genetic diseases such as primary hyperoxaluria (Danpure et al. 1993). A growing list of diseases caused by the improper localization of proteins makes protein translocation a promising target for the development of therapeutic agents (Besemer et al. 2005; Garrison et al. 2005).

Computational biologists have made extensive efforts to develop programs to predict the subcellular localization of proteins. Numerous software suites have been released in this field, based on various biological concepts and computational methods. Presently, four leading methods are commonly used. The first uses the overall protein amino acid composition. For example, SubLoc predicts protein localization based on the fact that proteins with different subcellular localizations usually have different amino acid compositions (Hua and Sun 2001). The second type of method utilizes known targeting sequences. One of the most important principles of the protein sorting mechanism is the existence of a targeting signal in the amino acid

sequence that leads proteins to different organelles or out of the cell. Hence, several computational approaches focus on predicting the presence of certain targeting motifs in protein sequences, e.g. signal peptides (SPs), the mitochondrial targeting peptide (mTP), nuclear localization signals (NLS) and transmembrane alpha helices (Bannai et al. 2002; Claros and Vincens 1996; Emanuelsson 2002). A third approach uses sequence homology and/or motifs. For example, the Proteome Analyst Subcellular Localization Server (PA-SUB) utilizes keywords from the protein database SWISS-PROT and the annotation of homologous proteins (Lu et al. 2004). Finally, a combination of the information obtained from the three categories described above has been used in prediction tools such as WoLF-PSORT (updated version of PSORT II) and the most recent, SherLoc2 (Horton et al. 2007; Briesemeister et al. 2009).

Due to their automated and high-throughput nature, computational methods are appealing for the large-scale assignment of protein subcellular locations. Regardless of the algorithm used, however, computational predictions have always been based on available biological knowledge, which is far from complete. The enormous complexity of the protein sorting process, the existence of alternative transportation pathways and the lack of complete data for every organelle still limit the application of computational methods. For instance, very few current predictors can deal with multi-site localization of a protein, with the exception of WoLF-PSORT and Hum-mPLoc (Shen and Chou 2009).

Due to the uncertain effectiveness of the available methods, particularly on a random protein dataset, we performed a comparative analysis between experimentally obtained subcellular localization data for 52 human Chr.21 proteins (Hu et al. 2006) and *in silico* prediction results, with the aim of evaluating the reliability of the bioinformatics approaches. Nine leading computational programs were included in the analysis, mainly due to their variable prediction strategies and the user-friendly web services that they provide.

Materials and methods

The materials and methods for the experimental characterization of protein subcellular localizations were reported previously (Hu et al. 2006). The computational predictions were performed on the internet website interfaces provided by each prediction program. A positive prediction was counted if the program gave the same site as at least one of the experimentally determined localizations for a given protein. The web addresses of the prediction programs used in this study are as follows: SherLoc2: http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc2; WoLF-PSORT: http://wolfpsort.org/; pTARGET: http://bioapps.rit.albany.edu/pTARGET/; ProtComp8: http://linux1.softberry.com/berry.phtml?topic=protcompan&group=programs&subgroup=proloc; PA-SUB v2.5: http://pasub.cs.ualberta.ca:8080/pa/Subcellular; MultiLoc2: http://www-bs.informatik.uni-tuebingen.de/Services/MultiLoc2/; ESLPred2: http://www.imtech.res.in/raghava/eslpred2/; BaCelLo: http://gpcr.biocomp.unibo.it/bacello/; SubLoc: http://www.bioinfo.tsinghua.edu.cn/SubLoc/.

Results

We divided the nine programs into two groups according to their prediction resolutions: low-resolution four-site prediction (nucleus, cytoplasm, mitochondrion and secretory pathway) and high-resolution organelle prediction that can further assign a secretory pathway protein to specific subcellular organelles such as the ER, Golgi apparatus, peroxisome and lysosome, as well as the plasma membrane and extracellular secretion. The prediction principles and capabilities of the nine programs are summarized in Table 1.

The prediction results for the 52 Chr.21 proteins are summarized in Tables 2, 3; they were compared to the experimentally determined localization patterns described previously (Hu et al. 2006). If one of the actual localization sites of a protein was predicted by a program, we counted a full positive prediction. This means, for example, that a prediction of "extracellular/secretory" in a low-resolution group was considered to reflect good performance in predicting the localization of plasma membrane, ER, Golgi and lysosomal proteins (in total, 15 proteins in this study). This loose criterion for the secretory pathway, however, was not applied to the high-resolution predictors that can classify proteins into specific organelle locations. For all of the predictors, however, a prediction of either "cytoplasm" or "nucleus" was counted as a full positive hit for the 12 Chr.21 proteins with "cyto-nuc" (cytoplasm and nucleus) dual localization. These calculations significantly raised the overall success rates for all nine of the predictors, but they should have no impact on comparisons of the relative performances of predictors with the same resolution, as none of the nine predictors showed a dual-localization prediction for any of the 52 proteins tested in this study.

The total number of positive predictions consistent with the experimental findings was summarized for each program; the percentage of prediction accuracy is shown next to the name of the prediction program in Figs. 1, 2. Among the low-resolution predictors, the three recently published programs MultiLoc2, ESLPred2 and BaCelLo were found to have similar prediction accuracies, with 75% (MultiLoc2-LowReso, ESLPred2) and 71% (BaCelLo) agreement

Table 1 Comparison of the protein localization prediction software programs used in the study

Software	Prediction strategy	Number of predicted localizations*	Reference
SherLoc2	Sequence-based predictions (aa composition, sorting signals), homology similarity, GO terms	9	Briesemeister et al. (2009)
WoLF-PSORT	Sequence-based predictions (aa composition, sorting signals, functional motifs), homology similarity	11	Horton et al. (2007)
pTARGET	Sequence-based predictions (aa composition, localization-specific Pfam domains)	9	Guda (2006)
ProtCom p8	Sequence-based predictions (signal sequences, anchors, other functional peptides), homology similarity	9	www.softberry.com
PA-SUB v.2.5	Homology similarity	9	Lu et al. (2004)
MultiLoc2[#]	Sequence-based predictions (aa composition, sorting signals), homology similarity, GO terms	4	Blum et al. (2009)
ESLPred2	Sequence-based predictions (aa composition, sorting signals), homology similarity	4	Garg and Raghava (2008)
BaCelLo	Sequence composition	4	Pierleoni et al. (2006)
SubLoc	Aa composition	4	Hua and Sun (2001)

* The number of sites was counted only for eukaryotic proteins

[#] Only the low-resolution function of MultiLoc2 was used; the high-resolution module was included in SherLoc2; *aa* amino acids

with the experimental data. A relatively low percentage of positive prediction, 60%, was observed for SubLoc, which was written in 2001.

The high-resolution predictors were found to have huge differences in accuracy. SherLoc2 and WoLF-PSORT displayed the highest accuracy, at 83 and 75%, respectively, which was significantly better than pTARGET (60%), ProtComp8 (56%) and PA-SUB v2.5 (54%). This variation in performance may originate from the different prediction methods that each program utilizes. There is a commonality among the two best predictors in both resolution groups (MultiLoc2 and ESLPred2, and SherLoc2 and WoLF-PSORT) in that they all utilize a wide range of prediction methods based on amino acid sequence composition, sorting signals and homology similarity. This finding indicates that the combination of homology information with sequence-based prediction can greatly improve the accuracy of protein localization prediction. On the other hand, the low success rate of PA-SUB (54%) suggested that searching for the localization of homologs alone is not powerful enough to create a high-standard prediction. The main problem of an approach based only on homology is that the prediction results can be ambiguous if there are no homologous proteins available with annotated localizations. In this study the localization of 10 out of 52 proteins could not be predicted using PA-SUB. This incompleteness creates a significant challenge when using homolog-based programs for genome-wide predictions of protein localization.

To evaluate whether prediction performance was associated with the specific localization site, the prediction results were grouped into different categories based on the experimental localization results. The number of predictions consistent with the experimental data was counted for each localization category and is shown in Figs. 1, 2. For the low-resolution predictors, the localization sites appeared to be irrelevant to prediction performance; the only exception was SubLoc, which could only predict seven out of 16 cytoplasmic proteins, a much smaller number than obtained with the other three programs. The performance similarity of these programs seemed reasonable because about 30% of the test proteins fell into the secretory pathway category.

When we looked at the data from the high-resolution predictors, the prediction accuracies were found to be closely correlated with the localization sites. For example, PA-SUB showed high accuracy in predicting cytoplasmic proteins (13 out of 16) but failed to predict all 12 of the plasma membrane proteins, of which over 80% could be predicted by the other four predictors. ProtComp8 and pTARGET, on the other hand, tended to have lower accuracy in predicting cytoplasmic proteins, scoring below 40%. A different trend was observed for the prediction of ER proteins. Interestingly, in spite of the existence of a signal peptide (SP)—the first and most extensively studied protein sorting signal—all five of the predictors tended to miss the proteins residing in the ER. Instead, the ER proteins (e.g., C21orf69 and TMPRSS3a) were often misclassified as extracellular secretory and plasma membrane proteins. This is very likely due to the biological fact that most secretory and plasma membrane proteins also carry an SP in their amino acid sequences.

Table 2 Comparison of experimental localization results for 52 Chr.21 proteins to *in silico* low-resolution predictions

Gene symbol	GenBank protein acc. no.	Function class	Localization in HEK293T	Low-resolution localization prediction			
				MultiLoc2-LowRes	ESLPred2	BaCelLo	SubLoc
ABCG1	CAA62631.1	ATPase	PM/Golgi	Cyto	Cyto	Cyto	Secr. Path.
AGPAT3	AAH11971.1	Acyltransferase	ER/PM(less)	Mito	Secr. Path	Secr. Path.	Cyto
B3GALT5	NP_006048.1	Galactosyl-transferase	Golgi/ER	Secr. Path.	Secr. Path.	Cyto	Mito
BACH1	BAA24932.1	Transcription regulation	Cyto(punct) Nuc-M-phase	Nuc	Nuc	Nuc	Nuc
C21orf103	NP_853633.1	Unclear	Cyto	Cyto	Secr. Path.	Secr. Path.	Nuc
C21orf19	AAL34462.1	Unknown	Nuc/Cyto	Cyto	Cyto	Cyto	Nuc
C21orf25	XP_032945.2	Unknown	Nuc/Cyto	Cyto	Nuc	Nuc	Nuc
C21orf30	CAB56001.2	Unknown	Nuc	Cyto	Nuc	Nuc	Nuc
C21orf4	AAC05974.2	Unknown	PM	Cyto	Secr. Path.	Secr. Path.	Cyto
C21orf59	AAG00496.1	Unknown	Nuc/Cyto	Cyto	Mito	Cyto	Cyto
C21orf69	AAK60445.1	Unknown	ER	Mito	Nuc	Secr. Path.	Nuc
C21orf96	NP_079419.1	Unknown	Cyto (punct)	Nuc	Nuc	Mito	Mito
CBS	NP_000062.1 (splicing isoform)	Cystathionine-beta-synthase	Cyto	Cyto	Cyto	Cyto	Cyto
CCT8	BAA02792.1	Chaperonin	Cyto	Cyto	Cyto	Cyto	Cyto
CHAF1B	NP_005432.1	Chromatin assembly factor	Nucleoplasm Cyto-M phase	Cyto	Nuc	Nuc	Cyto
CLDN14	AAG60052.1	Tight junction	ER/PM(less)	Secr. Path.	Secr. Path.	Secr. Path.	Secr. Path.
CLDN17	CAB60616.1	Tight junction	PM/Golgi	Secr. Path.	Secr. Path.	Secr. Path.	Secr. Path.
CLDN8	NP_036264.1	Tight junction	ER/PM(less)	Secr. Path.	Secr. Path.	Secr. Path.	Secr. Path.
CRYZL1	BAA91605.1	Oxidoreductase	Cyto	Cyto	Cyto	Cyto	Cyto
CXADR	AAH10536.1	Receptor	PM	Secr. Path.	Secr. Path.	Cyto	Nuc
DNMT3L	AAH02560.1	Methyltransferase-like	Nuc/Cyto	Cyto	Secr. Path.	Cyto	Cyto
DSCR3	NP_006043.1	Unknown	Nuc	Cyto	Cyto	Cyto	Cyto
ETS2	NP_005230.1	Transcription factor	Nuc	Nuc	Nuc	Cyto	Nuc
GCFC	AAD34617.1	Transcriptional repressor	Cyto	Nuc	Nuc	Nuc	Nuc
HLCS	NP_000402.2	Protein ligase	Cyto	Cyto	Nuc	Cyto	Cyto
HMGN1	AAA52676.1	DNA binding	Nuc	Nuc	Nuc	Nuc	Nuc
HSF2BP	NP_008962.1	Transcription factor binding	Cyto	Cyto	Nuc	Cyto	Cyto
IFNGR2	AAH03624.1	Receptor	ER/PM(less)	Secr. Path	Secr. Path.	Secr. Path.	Secr. Path.
KCNE1	AAH36452.1	K-channel	Lyso/PM	Secr. Path.	Secr. Path.	Cyto	Secr. Path.
KCNE2	NP_005127.1	K-channel	Lyso/PM	Cyto	Secr. Path.	Secr. Path.	Secr. Path.
KCNJ15	NP_002234.2	K-channel	PM/Golgi	Cyto	Cyto	Cyto	Cyto
KCNJ6	NP_002231.1	K-channel	PM/Golgi	Cyto	Cyto	Cyto	Cyto

Table 2 continued

Gene symbol	GenBank protein acc. no.	Function class	Localization in HEK293T	Low-resolution localization prediction			
				MultiLoc2-LowRes	ESLPred2	BaCelLo	SubLoc
KIAA0179	XP_035973.4	Unknown	Nuc/Cyto (punct)-M phase	Nuc	Nuc	Nuc	Nuc
MCM3AP	BAA25170.1	DNA binding	Cyto/Nuc	Cyto	Cyto	Cyto	Cyto
MX1	NP_002453.1	Dynamin and large GTPases	Cyto(punct)	Cyto	Mito	Cyto	Cyto
NNP1	AAH00380.1	RNA processing	Nucleolus	Nuc	Nuc	Nuc	Cyto
PCBP3	AAH12061.1	RNA binding	Cyto/Nuc	Cyto	Cyto	Nuc	Cyto
PCP4	CAA63724.1	Unknown	Nuc/Cyto	Cyto	Nuc	Cyto	Mito
PDE9A	AAH09047.1	Phosphodiesterase	Cyto (accum)	Cyto	Cyto	Nuc	Nuc
PDXK	AAH00123.1	Kinase	Cyto	Cyto	Cyto	Cyto	Secr. Path.
PFKL	AAH09919.1	Kinase	Cyto (accum)	Cyto	Cyto	Mito	Mito
PKNOX1	AAH07746.1	Transcription factor	Nuc/Cyto	Nuc	Nuc	Nuc	Nuc
PPIA3L	CAA37039.1	Peptidylprolyl isomerase A	Nuc/Cyto	Cyto	Cyto	Secr. Path.	Cyto
RPS5L	Pseudogene, 81% identity to BAB79493.1	Unknown	Cyto	Cyto	Cyto	Cyto	Secr. Path.
SH3BGR	AAH06371.1	SH3 adaptor	Cyto	Cyto	Cyto	Cyto	Nuc
TAK1L	AAF81754.1	Transcription factor-like	Nuc/Cyto	Cyto	Nuc	Cyto	Secr. Path.
TMPRSS3a	NP_076927.1	Protease	ER	Cyto	Secr. Path.	Cyto	Secr. Path.
TSGA2	NP_543136.1	Chromosome-associated	Cyto/Nuc	Cyto	Cyto	Cyto	Cyto
UBASH3A	NP_061834.1	Catalytic activity	Cyto	Cyto	Cyto	Cyto	Cyto
UBE2G2	AAC32312.1	Ligase	Cyto	Cyto	Cyto	Cyto	Nuc
WDR4	AAH06341.1	Unknown	Nucleoplasm	Secr. Path	Nuc	Nuc	Cyto
WDR9_3′	BAA92123.1	Unknown	Nuc	Nuc	Nuc	Nuc	Nuc

The localization properties of 52 Chr.21 proteins determined experimentally in HEK293T cells were compared to prediction results given by four computational programs that can only classify proteins into four subcellular compartments. *Accum* accumulated, *Cyto* cytosol, *ER* endoplasmic reticulum, *Lyso* lysosome and endosome, *Mem-bound* membrane-bound, *Mito* mitochondria, *Nuc* Nucleus, *PM* plasma membrane, *Punct* punctuated, *Secr. Path.* extracellular secreted protein or secretory pathway protein

Table 3 Comparison of experimental localization results for 52 Chr.21 proteins to *in silico* high-resolution predictions

Gene symbol	GenBank protein acc. no.	Function class	Localization in HEK293T	High-resolution localization prediction				
				SherLoc2	WoLF-PSORT	pTARGET	ProtComp8	PA-SUB v2.5
ABCG1	CAA62631.1	ATPase	PM/Golgi	Cyto	PM	PM	PM	ER
AGPAT3	AAH11971.1	Acyltransferase	ER/PM(less)	ER	Extracell	ER	ER	Mito
B3GALT5	NP_006048.1	Galactosyl-transferase	Golgi/ER	Golgi	Extracell	Golgi	Golgi	Golgi
BACH1	BAA24932.1	Transcription regulation	Cyto(punct) Nuc-M-phase	Nuc	Nuc	Nuc	Nuc	Nuc
C21orf103	NP_853633.1	Unclear	Cyto	Cyto	Extracell	Extracell	PM	Cyto
C21orf19	AAL34462.1	Unknown	Nuc/Cyto	Mito	Nuc	PM	Extracell	–
C21orf25	XP_032945.2	Unknown	Nuc/Cyto	Nuc	Nuc	Nuc	Extracell	Extracell
C21orf30	CAB56001.2	Unknown	Nuc	Cyto	Nuc	Extracell	Mem-bound Perox	–
C21orf4	AAC05974.2	Unknown	PM	PM	PM	PM	Extracell	–
C21orf59	AAG00496.1	Unknown	Nuc/Cyto	Cyto	Cyto	Cyto	Extracell	–
C21orf69	AAK60445.1	Unknown	ER	Mito	Extracell	Extracell	Cyto	–
C21orf96	NP_079419.1	Unknown	Cyto (punct)	Cyto	Cyto_Nuc	Cyto	Extracell	–
CBS	NP_000062.1 (splicing isoform)	Cystathionine-beta-synthase	Cyto	Cyto	PM	–	Cyto	Cyto
CCT8	BAA02792.1	Chaperonin	Cyto	Cyto	Cyto	Mito	Cyto	Cyto
CHAF1B	NP_005432.1	Chromatin assembly factor	Nucleoplasm Cyto-M phase	Nuc	Nuc	Nuc	Nuc	Nuc
CLDN14	AAG60052.1	Tight junction	ER/PM(less)	PM	PM	PM	PM	–
CLDN17	CAB60616.1	Tight junction	PM/Golgi	PM	PM	PM	PM	–
CLDN8	NP_036264.1	Tight junction	ER/PM(less)	PM	PM	PM	PM	–
CRYZL1	BAA91605.1	Oxidoreductase	Cyto	Cyto	Cyto	Cyto	Extracell	Cyto
CXADR	AAH10536.1	Receptor	PM	PM	PM	Extracell	PM	Extracell
DNMT3L	AAH02560.1	Methyltransferase-like	Nuc/Cyto	Cyto	Nuc	PM	Nuc	Nuc
DSCR3	NP_006043.1	Unknown	Nuc	Cyto	Cyto	Cyto	Extracell	Cyto
ETS2	NP_005230.1	Transcription factor	Nuc	Nuc	Nuc	Nuc	Nuc	Nuc
GCFC	AAD34617.1	Transcriptional repressor	Cyto	Nuc	Cyto	Nuc	Mem-bound perox	Nuc
HLCS	NP_000402.2	Protein ligase	Cyto	Cyto	Cyto	Mito	Extracell	Cyto
HMGN1	AAA52676.1	DNA binding	Nuc	Nuc	Nuc	Nuc	Mito	Nuc
HSF2BP	NP_008962.1	Transcription factor binding	Cyto	Cyto	Cyto	Cyto	Extracell	Cyto
IFNGR2	AAH03624.1	Receptor	ER/PM (less)	PM	PM	Lyso	PM	Extracell
KCNE1	AAH36452.1	K-channel	Lyso/PM	PM	Extracell	PM	PM	ER
KCNE2	NP_005127.1	K-channel	Lyso/PM	PM	Cyto	PM	PM	ER
KCNJ15	NP_002234.2	K-channel	PM/Golgi	PM	PM	PM	PM	ER
KCNJ6	NP_002231.1	K-channel	PM/Golgi	PM	PM	PM	PM	Mito
KIAA0179	XP_035973.4	Unknown	Nuc/Cyto (punct)-M phase	Nuc	Nuc	Nuc	Nuc	Nuc

Table 3 continued

Gene symbol	GenBank protein acc. no.	Function class	Localization in HEK293T	High-resolution localization prediction				
				SherLoc2	WoLF-PSORT	pTARGET	ProtComp8	PA-SUB v2.5
MCM3AP	BAA25170.1	DNA binding	Cyto/Nuc	Nuc	Nuc	Lyso	Extracell	Nuc
MX1	NP_002453.1	Dynamin and large GTPases	Cyto (punct)	Cyto	Cyto	Cyto	Cyto	Cyto
NNP1	AAH00380.1	RNA processing	Nucleolus	Nuc	Nuc	Nuc	Nuc	Nuc
PCBP3	AAH12061.1	RNA binding	Cyto/Nuc	Nuc	Cysk	Cyto	Cyto	Cyto
PCP4	CAA63724.1	Unknown	Nuc/Cyto	Cyto	Cyto	Cyto	Mito	Cyto
PDE9A	AAH09047.1	Phosphodiesterase	Cyto (accum)	Cyto	Cyto	Cyto	Extracell	Cyto
PDXK	AAH00123.1	Kinase	Cyto	Cyto	Cyto	Cyto	Cyto	Cyto
PFKL	AAH09919.1	Kinase	Cyto (accum)	Cyto	Cyto	–	Cyto	Cyto
PKNOX1	AAH07746.1	Transcription factor	Nuc/Cyto	Nuc	Nuc	Cyto	Nuc	Nuc
PPIA3L	CAA37039.1	Peptidylprolyl isomerase A	Nuc/Cyto	Cyto	Cyto	Cyto	Cyto	Cyto
RPS5L	Pseudogene, 81% identity to BAB79493.1	Unknown	Cyto	Cyto	Cyto	–	Extracell	–
SH3BGR	AAH06371.1	SH3 adaptor	Cyto	Cyto	Cyto	Nuc	Extracell	Cyto
TAK1L	AAF81754.1	Transcription factor-like	Nuc/Cyto	Cyto	Cyto	Cyto	Cyto	Cyto
TMPRSS3a	NP_076927.1	Protease	ER	PM	Cyto	PM	ER	Extracell
TSGA2	NP_543136.1	Chromosome-associated	Cyto/Nuc	Cyto	Cyto	Cyto	Extracell	Cyto
UBASH3A	NP_061834.1	Catalytic activity	Cyto	Cyto	Nuc	Golgi	Extracell	Cyto
UBE2G2	AAC32312.1	Ligase	Cyto	Perox	Mito	ER	Extracell	Cyto
WDR4	AAH06341.1	Unknown	Nucleoplasm	Cyto	Extracell	Golgi	Extracell	Cyto
WDR9_3′	BAA92123.1	Unknown	Nuc	Nuc	Nuc	–	Nuc	Cyto

The localization properties of 52 Chr.21 proteins determined experimentally in HEK293T cells were compared to prediction results given by five computational programs that can classify proteins into at least nine subcellular compartments. *Accum* accumulated *Cysk* cytoskeleton, *Cyto* cytosol, *ER* endoplasmic reticulum, *Extracell* extracellular secreted protein, *Lyso* lysosome and endosome, *Mem-bound* membrane-bound, *Mito* mitochondria, *Nuc* Nucleus, *PM* plasma membrane, *Punct* punctuated, *Perox* peroxisome

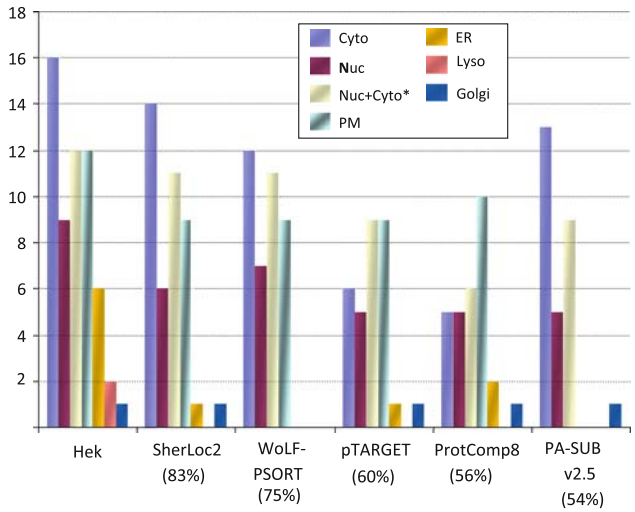

Fig. 1 Comparison of the prediction performances of five computational predictors with high resolution. Prediction performance varied among the different programs. SherLoc2 and WoLF-PSORT rendered the highest accuracy with the experimental results (indicated as Hek), at 83% and 75%, respectively, which was significantly better than pTARGET (60%), ProtComp8 (56%) and PA-SUB v2.5 (54%). Prediction accuracy was found to be associated with the specific localization site. Abbreviations: *Nuc* nucleus, *Cyto* cytoplasm, *PM* plasma membrane, *ER* endoplasmic reticulum, *Lyso* lysosome and endosome. *For the proteins with dual localization sites, all five of the predictors predicted only one site but such predictions were still counted as a full correct prediction

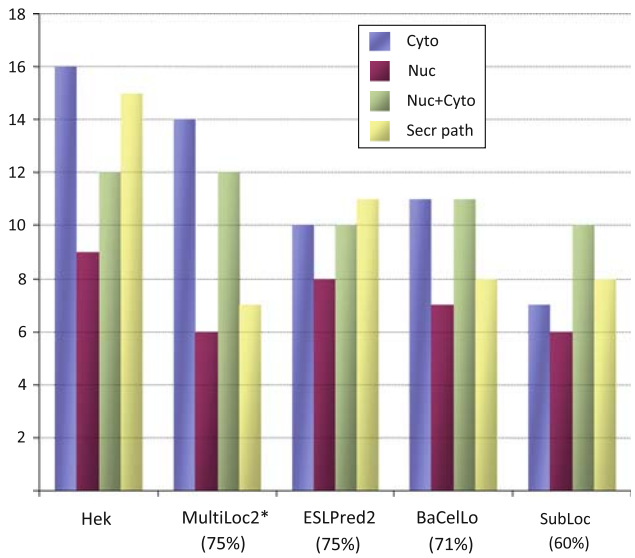

Fig. 2 Comparison of the prediction performances of four computational predictors with low resolution. The recently developed predictors were found to have similar prediction accuracies, with 75% (MultiLoc2-LowReso, ESLPred2) and 71% (BaCelLo) agreement with the experimental data (indicated as Hek). A relatively low percentage of positive prediction, 60%, was observed for SubLoc, which was developed in 2001. Prediction accuracy was found to be associated with the specific localization site. Abbreviations: *Nuc* nucleus, *Cyto* cytoplasm, *Secr. path.* secretory pathway protein (including plasma membrane, ER, Golgi and lysosomal proteins in this study)

Discussion

The localization site-dependent performance shown by the different prediction programs may be attributable to the different prediction strategies utilized by each particular program and the level of knowledge available about protein trafficking mechanisms. For example, the sequence and structure of the signal peptide (SP), a motif that directs proteins to the ER membrane, are well studied as compared to nuclear localization signals (NLS), thus facilitating the prediction of proteins destined for the ER-associated secretory pathway (e.g., ER, Golgi, plasma membrane, lysosome/endosome and secretory proteins). This contributes to the high accuracy of low-resolution predictors that do not distinguish between specific localization sites within the pathway. For the high-resolution predictors, however, difficulties remain regarding how to classify the different organelles in relation to the secretory pathway. Hence, further studies on protein targeting motifs and their underlying mechanisms should contribute to the improvement of the accuracy of protein localization predictions.

The present results demonstrate that prediction performance varies between different programs and different localization categories. Consequently, it might be advisable to use multiple localization predictors that utilize different prediction methods. Moreover, special attention should be paid to the relative confidence scores assigned to the different localization sites. Generally, a large difference between the second best score and the best one implies a reliable prediction, whereas similar scores obtained for different locations may reflect the unreliability of the prediction or may indicate that the protein has multiple localization patterns. A good example of this in our study is the C21orf7 protein. The *C21orf7* (TAK1-like) gene shares homology with the human *TAK1* (TGF-beta activated kinase) gene, which plays a critical role in the TGF-beta signal transduction pathway. Even though it was classified as a cytoplasmic protein by most of the predictors, ESLPred2 predicted the nucleus as the most plausible localization site; moreover, WoLF-PSORT suggested a dual localization in the cytoplasm and nucleus with 19.8% probability, second to a 24% probability of localization in the cytoplasm alone. In our previous transfected-cell array experiments (Hu et al. 2006), the actual localization of this protein was found to be quite dynamic, with a distribution in both the cytoplasm and the nucleus.

In some cases the predictions may still be incorrect even though the majority of the predictors report the same localization. In this study the actual localization of several proteins was in disagreement with most of the predictions. For example, the *WDR4* gene encodes a member of the WD-repeat protein family and is a candidate for some disorders mapped to 21q22.3 and for Down syndrome phenotypes (Michaud et al. 2000). Despite the fact that

BaCelLo and ESLPred2 predicted it as a nuclear protein, the other seven programs predicted that it is either cytoplasmic protein or is exported outside of the cell. In the actual experiment, WDR4 proteins were found to reside in the nucleus, distributed within the nucleoplasm. The yeast homolog of WDR4, Trm82, has been previously reported to be required for 7-methylguanosine modification of tRNA (Alexandrov et al. 2002). Because this pre-tRNA processing is known to take place in the nucleoplasm before the resulting mature tRNAs are transported out to the cytoplasm (Lodish et al. 2000), Trm82 was expected to localize in the nucleus, especially in the nucleoplasm, as we observed for WDR4. Although the functional role of WDR4 in human cells has not been experimentally verified, Alexandrov et al. have found that WDR4, in a complex with METTL1, is required for the 7-methylguanosine modification of yeast tRNA (Alexandrov et al. 2002). In conjunction with our localization results, this finding suggests that human WDR4 performs a similar tRNA-processing function as does its yeast homolog.

Taken together, despite the relatively small number of proteins analyzed in this study, our results indicate a generally lower percentage of prediction accuracy (54–83%) than claimed by recently published predictors; for instance, ESLPred2 was claimed to have an accuracy of over 90% (Garg and Raghava 2008). Nevertheless, SherLoc2, MultiLoc2, ESLPred2 and WoLF-PSORT showed significantly better performance than the other programs evaluated in our study. The predictors that showed the best performance were SherLoc2 and WoLF-PSORT. Both programs can carry out high-resolution predictions of at least nine subcellular localizations, which is an extra merit in addition to their high prediction accuracy. Their outstanding capabilities are likely related to the multi-dimensional biological information they integrate into their prediction strategies, ranging from amino acid composition and the presence of sorting signals and targeting motifs to homology profiles and Gene Ontology terms.

Taken together, the differences in the accuracy of subcellular protein localization predictions presented in this study strongly suggest that the outcomes of *in silico* localization predictions should be treated with caution, and that it is always beneficial to compare the results provided by different prediction algorithms.

Acknowledgments We thank Sebastian Briesemeister from Eberhard Karls University Tübingen for helping with the predictions using SherLoc2 and MultiLoc2. This study was supported by the Max Planck Society.

References

Alexandrov A, Martzen MR, Phizicky EM (2002) Two proteins that form a complex are required for 7-methylguanosine modification of yeast tRNA. RNA 8:1253–1266

Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S (2002) Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 18:298–305

Besemer J, Harant H, Wang S, Oberhauser B, Marquardt K, Foster CA, Schreiner EP, de Vries JE, Dascher-Nadel C, Lindley IJ (2005) Selective inhibition of cotranslational translocation of vascular cell adhesion molecule 1. Nature 436:290–293

Blum T, Briesemeister S, Kohlbacher O (2009) MultiLoc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. BMC Bioinformatics 10:274

Briesemeister S, Blum T, Brady S, Lam Y, Kohlbacher O, Shatkay H (2009) SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. J Proteome Res 8:5363–5366

Claros MG, Vincens P (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem 241:779–786

Danpure CJ, Purdue PE, Fryer P, Griffiths S, Allsop J, Lumb MJ, Guttridge KM, Jennings PR, Scheinman JI, Mauer SM et al (1993) Enzymological and mutational analysis of a complex primary hyperoxaluria type 1 phenotype involving alanine: glyoxylate aminotransferase peroxisome-to-mitochondrion mistargeting and intraperoxisomal aggregation. Am J Hum Genet 53:417–432

Emanuelsson O (2002) Predicting protein subcellular localisation from amino acid sequence information. Brief Bioinform 3:361–376

Garg A, Raghava GPS (2008) ESLpred2: Improved method for predicting subcellular localization of eukaryotic proteins. BMC Bioinformatics 9:503

Garrison JL, Kunkel EJ, Hegde RS, Taunton J (2005) A substrate-specific inhibitor of protein translocation into the endoplasmic reticulum. Nature 436:285–289

Guda C (2006) pTARGET: a web server for predicting protein subcellular localization. Nucleic Acids Res 35:W210–W213

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) "WoLF PSORT: Protein Localization Predictor". Nucleic Acids Res 35:W585–W587 (Web Server issue)

Hu YH, Warnatz HJ, Vanhecke D, Wagner F, Fiebitz A, Thamm S, Kahlem P, Lehrach H, Yaspo ML, Janitz M (2006) Cell array-based intracellular localization screening reveals novel functional features of human chromosome 21 proteins. BMC Genomics 7:155

Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17:721–728

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore P, Darnell JE (2000) Molecular Cell Biology. Freeman WH & Co

Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C, Eisner R (2004) Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics 20:547–556

Michaud J, Kudoh J, Berry A, Bonne-Tamir B, Lalioti MD, Rossier C, Shibuya K, Kawasaki K, Asakawa S, Minoshima S, Shimizu N, Antonarakis SE, Scott HS (2000) Isolation and characterization of a human chromosome 21q22.3 gene (WDR4) and its mouse homologue that code for a WD-repeat protein. Genomics 68:71–79

Pierleoni A, Martell PL, Farisell P, Casadio R (2006) BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22:e408–e416

Relic B, Andjelkovic M, Rossi L, Nagamine Y, Hohn B (1998) Interaction of the DNA modifying proteins VirD1 and VirD2 of Agrobacterium tumefaciens: analysis by subcellular localization in mammalian cells. Proc Natl Acad Sci USA 95:9105–9110

Shen HB, Chou KC (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. Anal Biochem 394:269–274

Surapureddi S, Svartz J, Magnusson KE, Hammarstrom S, Soderstrom M (2000) Colocalization of leukotriene C synthase and microsomal glutathione Stransferase elucidated by indirect immunofluorescence analysis. FEBS Lett 480:239–243

Zuckerbraun BS, Shapiro RA, Billiar TR, Tzeng E (2003) RhoA influences the nuclear localization of extracellular signal-regulated kinases to modulate p21Waf/Cip1 expression. Circulation 108:876–881